DATA

Zack Kertcher

Statistics for Management
Fall 2016

Plan for today

1. Data structure and data types

2. *Getting data

DATA STRUCTURES AND TYPES

What types of data exist?

Nominal (factor)	Interval (numeric)	Nominal (factor)	Nominal (factor)	Interval (numeric)	Ordinal (numeric/factor)	Nominal (factor)	Nominal (character)
state	account_len	intl_plan	vm_plan	min_day	satisfy	churn	comments
HI	101	no	no	70.9	4	no	###########
MT	137	no	no	223.6	3	yes	###########
OH	103	no	yes	294.7	4	no	###########
NM	99	no	no	216.8	1	yes	###########
SC	108	no	no	197.4	2	no	###########
IA	117	no	no	226.5	3	no	###########
ND	63	no	yes	218.9	4	no	###########
LA	94	no	no	157.5	4	yes	###########
MO	138	no	no	89.1	2	no	###########

Cases
(rows)

Variables
(columns)

- ✓ Describe the data types found in this table

- ✓ Can you think of other ways to represent some of these data types?

- ✓ What would you need to do to conduct statistical analysis of these data?

title	year	length	budget	rating	votes	Action	Animation	Comedy	Drama	Documentary
3	1971	86	NA	6	14	0	0	0	1	0
Ikinai	1998	101		7.1	102	0	0	0	0	0
Rundown, The	2003	104	85000000	6.5	8067	1	0	1	0	0
	2001	101	NA	6.7	17		0	1	0	0
Somersault	2004	106		6.9	370	0	0	0	1	0
In Search of the Obelisk	1993	4	NA	5.8	8	0		0	0	0
Megaville	1990	88	NA	4.6	60	0	0	0	0	0

How are data organized?

Vector

Vector

Vector is the most common data structure in R (and other software)

Formally, it is a unidimensional sequence of data elements

A vector has the same data type

Common types: numeric, integer, character, logical

Vector

Examine a vector

`class()`	←	type of an object (e.g. logical, character)
`str()`	←	structure of an object
`is.na()`	←	expression to check for missing elements
`is.null()`	←	expression to check for empty vector
`length()`	←	length of elements in a vector
`nchar()`	←	number of characters per element

- ✓ Create a **statmngr** vector, with the following values: **9, 8, 9, 8.2, 7.1**
- ✓ What type of vector is it?
- ✓ Compute the mean of this vector, hint: use the **mean** function
- ✓ Add NA, 81, 7 at the end of **statmngr**
- ✓ Compute the mean of the new **statmngr**. What did you get?

Index

Index (Access data elements)

🟦🟩🟦	`vec[2]`	←	Second element
🟦🟩🟩	`vec[2:3]`	←	Second to third elements
🟩🟦🟩	`vec[c(1,3)]`	←	First and third elements
🟦🟩🟦	`vec[-c(1,3)]`	←	Exclude first and third elements

Rearrange vector

```
sort|order()

order(vec, decreasing=F)
```

Logical operators in R

Operator	Means
<	Less than
<=	Less than or equal to
>	More than
>=	More than or equal to
==	Equals to
!=	Not equals to
!x	Not x
x | y	x or y
x & y	x and y

- Create a **numvec** vector, with the values: 1**,3,5, 7….99, .** Hint: **seq( )**

- How many values are in this vector?

- From this vector, select all the numbers larger than 50 and assign them into a **morethan50** vector. Hint: format for this logical expression is **x > y**

- Reverse sort this vector. Save the 12,8,20th values as a **newnums** vector in this sequence. What numbers does **newnums** contain?

Change vector elements

By inputting new data (we already did this!)

By substituting and/or removing data

✓ Generate a **num** vector with the following numbers: (9,11, 13….99)

✓ What are the 2nd and 4th quantiles of **num**? Hint: **quantile( )**

✓ What is the standard deviation of **num**, when selecting from it numbers greater than 20? (e.g., 21,24…99) Hint: **sd( )**

🏆 What is the mean of numbers from **num** that are less than 51 and more than 71?

Change vector type

test for data type

```
is.character|factor|numeric|integer|logical…()

as.character|logical|numeric|integer|factor…()
```

Coerce data type

- ✓ Sort the **statmngr** you created earlier from the smallest number
- ✓ Replace the second value with 7.9
- ✓ Replace 9 with an NA
- ✓ Has the class of the **statmngr** vector changed?
- ✓ What happens when you try to coerce this vector into a factor?

Data frame

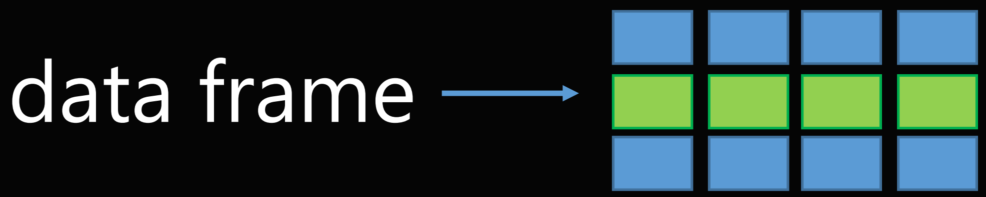
Data frame

Most common data type in R
(Yes, there are more, including matrix, array, and lists)

Two-dimensional

Accepts vectors of different vector types

Vector lengths need to be identical

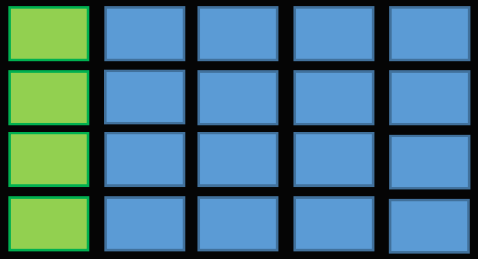


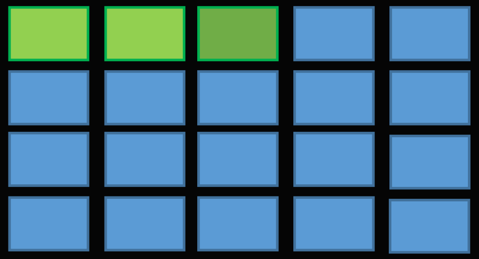
Create a data frame

By inputting data

By combining data from elsewhere, for example, by combining vectors

Data frame index (access elements)

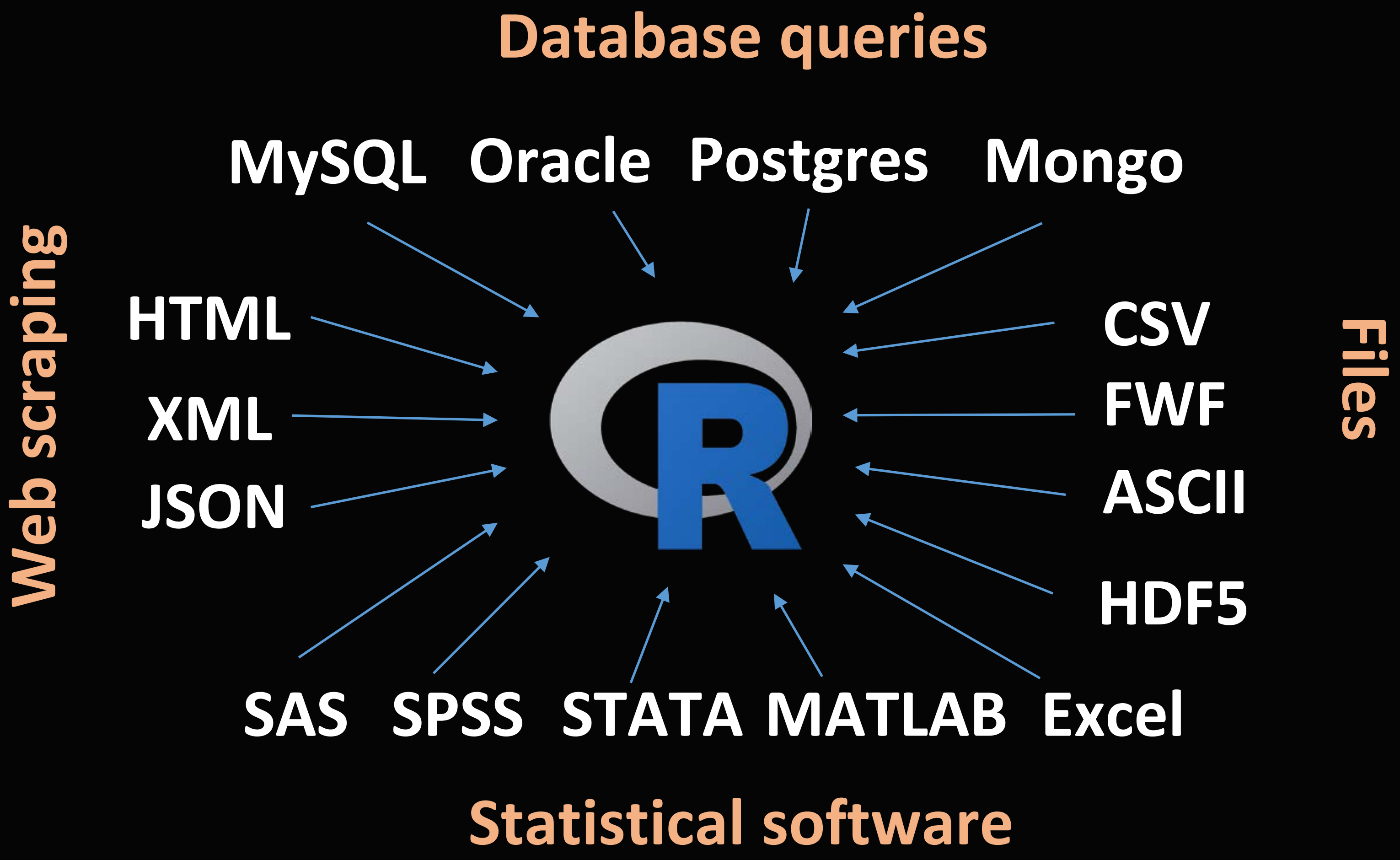




- ✓ Create a data frame **hr** with 2 columns: **Experience** (1,2,1,8) and **Performance** (10,8,7,10)
- ✓ Add a **Employees** column (3,4,2,10)
- ✓ Now add another row with **experience** 3, **performance** 7, and **employees** 4
- ✓ Add an **ID** column (12,13,14,15)
- ✓ Change the **Performance** of **ID** 14 to 9
- ✓ What is the average **Performance** in the **hr** data?

GETTING DATA

How do we get data?



Database queries

MySQL

Oracle

Postgres

Mongo

Web scraping

HTML

XML

JSON

Files

CSV

FWF

ASCII

HDF5

SAS

SPSS

STATA

MATLAB

Excel

Statistical software

Compressed files

Many other…

In this class we will mostly

Generate data (like we did so far) – useful for testing and simulation

Data from an online source – Kaggle, github, Chicago Data Portal, etc.

Various R packages (vincentarelbundock.github.io/Rdatasets/datasets.html)

Reading data from a file

Function	Used for reading
readLines	raw text files
read.csv	csv files separated by a ","
read.csv2	csv files separated by a ";"
read.delim	files separated by "\t"
fead.fwf	fixed format files
read.table	all the above plus more

Function	Used for reading	Package
read.xls	Excel spreadsheet (specific sheet)	xlsx (various others, like readxl)
read.spss	SPSS .sav files	foreign
read.dta	Stata .dta files	foreign
fead.xport	SAS .xpt files	foreign
read_sas	SAS .b7dat, b7cat	haven

✓ Read the retail data from **Blackboard** (Course Documents→Lecture Notes→Session 2→retail_samp.csv. Name the data frame **retail**

✓ Inspect the data using the various functions we learned so far.

✓ What are the variable types in the data?

✓ Note at least 3 things that you think you would need to do to make the data ready for analysis.